Abstract Title Page
Not included in page count.

Title: Evaluating Non-Randomized Educational Interventions: A Graphical Discussion

Authors and Affiliations:

Roddy Theobald, Department of Statistics, University of Washington
Thomas Richardson, Department of Statistics, University of Washington

Abstract Body

Background / Context:
A central goal of the education literature is to demonstrate that specific educational interventions–instructional interventions at the student or classroom level, structural interventions at the school level, or funding interventions at the school district level, for example–have a "treatment effect" on student achievement. In rare cases (e.g., Nye et al., 2000), researchers have the opportunity and resources to conduct randomized experiments in which the intervention or treatment is randomly assigned to units. However, randomization is often not feasible in educational settings. For example, it is not possible for a school to randomly assign federally-mandated interventions (like special education) to students, for a district to deprive resources from a random subset of schools, or for a state to randomly assign districts to one of two funding systems. Thus in the vast majority of cases, researchers must estimate the treatment effect of an educational intervention from observational data in which the intervention has not been randomly assigned to units.

Extensive literatures on estimating treatment effects from observational data exist in Economics (e.g., Imbens and Wooldridge, 2008), Sociology (e.g., Winship and Morgan, 1999), and Statistics (e.g., Rosenbaum and Rubin, 1983). The unifying concern is that estimates of treatment effects from observational data may be biased by confounding if there are unobserved variables that are correlated both with the probability of receiving the intervention and with the outcome of interest (in our case, student achievement). Countless papers have discussed the assumptions necessary for various estimation methods to produce unbiased estimates of a treatment effect in the presence of confounding.

However, researchers in different fields discuss these assumptions differently. Researchers in Statistics, Economics, and other social sciences often discuss confounding and causality with potential outcomes, while computer scientists, epidemiologists, and sociologists often utilize directed acyclic graphs (DAGs). Recently, Richardson and Robins (2013) introduced Single World Intervention Templates (SWITs), a unification of these approaches to causality. We argue that SWITs provide an intuitive graphical framework for education researchers to discuss assumptions related to confounding and causality.

Purpose / Objective / Research Question / Focus of Study:
This paper has three objectives. First, we explain both how SWITs unify two existing approaches to causality—potential outcomes and DAGs—and their utility as a framework for discussing confounding and causality. Second, we introduce a specific SWIT that can be used to discuss confounding and causality in a "pre/post" study of an educational intervention. Finally, we use this SWIT to discuss the assumptions necessary for common estimation methods to produce an unbiased estimate of the treatment effect of a non-randomized educational intervention.

Significance / Novelty of study:
For simplicity, we will discuss a typical "pre/post" observational study of an educational intervention. For each student in the sample, we observe test scores Y_{t-1} at a time *t-1* prior to the intervention and Y_t at a time *t* after the intervention. A subset of students receives an intervention in year *t* that is intended to have an effect on student performance in that year. For each student,

let $A_t = 1$ if the student receives the intervention in year t and $A_t = 0$ if the student does not receive the intervention in year t. The goal is to estimate the treatment effect of A_t on Y_t.

Potential outcomes provide one framework for discussing causality in this setting. Specifically, we can assume that each student in the study has two "potential outcomes": $Y_t(a_t = 1) = Y_t(1)$ is the student's test score in year *t* if (potentially contrary to fact) the student received the intervention; and $Y_t(a_t = 0) = Y_t(0)$ the student's test score in year *t* if the student did not receive the intervention. One potential estimand of interest is the "average treatment effect" $E\{Y_t(1) - Y_t(0)\}$; that is, the average difference between how students would have performed had they received the intervention and how students would have performed had they not received the intervention. In the potential outcomes literature, the effect of A_t on Y_t is unconfounded if $A_t \perp\!\!\!\perp Y_t(a_t)\ \forall\ a_t$. One common assumption in observational studies is that observed variables are sufficient to control for confounding between A_t and Y_t. In the potential outcomes literature, if X is an observed variable that may be correlated with both A_t and Y_t, the effect of A_t on Y_t is unconfounded conditional on X if $A_t \perp\!\!\!\perp Y_t(a_t) \mid X\ \forall\ a_t$.

Another framework for reasoning about confounding and causality is directed acyclic graphs (DAGs). Figure 1 shows two simple DAGs. Figure 1a shows a simple unconfounded relationship between A_t and Y_t, while figure 1b shows the same relationship that is confounded by an unobserved variable M. DAGs provide a simple graphical representation of confounding: if there is a "backdoor path" from A_t to Y_t (for example, the path $A_t \leftarrow M \rightarrow Y_t$ in figure 1b), then the effect of A_t on Y_t is confounded.

(Please insert Figure 1 here)

DAGs also provide a straightforward way to establish conditional independence relationships between variables. Suppose we observe a variable X about each student, and assume that the null hypothesis that A_t has not effect on Y_t holds. The DAG literature contains a criterion called d-separation (Pearl 2000) that allows us to establish that $A_t \perp\!\!\!\perp Y_t \mid X$. Table 1 contains the relevant definitions and theorems from this literature.

(Please insert Table 1 here)

Figure 2 illustrates two scenarios. In figure 2a, there are no paths d-connecting A_t and Y_t given X, so $A_t \perp\!\!\!\perp Y_t \mid X$ in figure 2a. Figure 2b, on the other hand, contains the path $A_t \leftarrow M \rightarrow Y_t$, which d-connects A_t and Y_t given X. Thus $A_t \not\perp\!\!\!\perp Y_t \mid X$ in figure 2b.

(Please insert Figure 2 here)

But DAGs do not allow us to relate these conditional independencies to the potential outcomes definition of confounding. For this reason, Richardson and Robins (2013) introduced Single World Intervention Templates (SWITs). Figure 3 shows the causal relationships in Figure 1 as SWITs. The vertex that represents the intervention in figure 3a is "split" into the random variable A_t and hypothetical intervention a_t, while the vertex representing the outcome Y_t has been

relabeled as a potential outcome that depends on the hypothetical intervention a_t.

(Please insert Figure 3 here)

SWITs give a clear graphical interpretation of confounding that is consistent with the potential outcomes definition. In figure 3a, there is no path from A_t to $Y_t(a_t)$, so $A_t \perp Y_t(a_t) \; \forall \; a_t$ and the effect of A_t on Y_t is unconfounded. In figure 3b, the backdoor path $A_t \leftarrow M \rightarrow Y_t(a_t)$ means that $A_t \not\perp Y_t(a_t) \; \forall \; a_t$ and the effect of A_t on Y_t is confounded.

More importantly, SWITs allow us to connect the potential outcomes definition of conditional confounding to the DAG criterion for conditional independence (d-separation, Table 1). Figure 4 contains SWITs representing the same two scenarios shown in Figure 2.

(Please insert Figure 4 here)

In figure 4a, the d-separation of A_t and $Y_t(a_t)$ conditional on X allows us to conclude that $A_t \perp Y_t(a_t) \mid X \; \forall \; a_t$ so the effect of A_t on Y_t is unconfounded conditional on X by the potential outcomes definition of confounding. In figure 4b, on the other hand, A_t and $Y_t(a_t)$ are d-connected conditional on X, so $A_t \not\perp Y_t(a_t) \mid X \; \forall \; a_t$ so the effect of A_t on Y_t is confounded conditional on X.

Statistical, Measurement, or Econometric Model:
In estimating the treatment effect of a non-randomized educational intervention A_t on student tests scores Y_t, researchers must worry about potential confounders that are observed and unobserved, as well as potential confounders that vary over time (time-variant) and do not vary over time (time-invariant). Researchers must also account for the reality that test scores can be noisy measures of student achievement. Table 2 summarizes the notation we use in this more realistic framework.

(Please insert Table 2 here)

Figure 5 shows the relationships between these variables in this framework. A key departure from the previous discussion is that student test scores Y_t are now an unbiased estimate of unobserved student achievement Y_t^*. In figure 5, black vertices represent variables that are observed to the researcher, while gray vertices represent unobserved variables. Black solid edges represent causal relationships between potential confounders and student achievement and test scores that are assumed to always exist. The red dotted line is the causal relationship of interest. Finally, the blue dotted lines represent causal relationships between potential confounders and the probability of treatment that may or may not exist depending on assumptions. A central message of our paper is that the assumptions that justify many common methods of evaluating educational interventions, such as covariate-adjustment models, student fixed effects models, and instrumental variables models, can be expressed in terms of the existence of and relationships between the blue dotted lines in figure 5.

(Please insert Figure 5 here)

Usefulness / Applicability of Method:
We will focus exclusively on covariate-adjustment models (e.g., regression or matching), although in our paper we also use the SWIT in figure 5 to discuss the assumptions that justify student fixed effects models and instrumental variables models. The assumption that justifies a covariate-adjustment model is the oft-cited "strong ignorability condition" (Rosenbaum and Rubin 1983). In terms of the variables in Table 2, this assumption can be written as follows:

Assumption 1: $A_t \perp U_{t-1}, Y^*_{t-1}, M \mid X, L_{t-1}, Y_{t-1}$.

In other words, the probability of receiving the intervention must be independent of all unobserved variables, conditional on the observed variables. Figure 6 shows the causal relationships under assumption 1. The black dashed edges in figure 6 represent relationships between potential confounders and the probability of receiving the intervention that are allowed to exist under this assumption. In figure 6, there is no path that d-connects A_t and $Y^*_t(a_t)$ given X, L_{t-1}, and Y_{t-1}, so $A_t \not\perp Y^*_t(a_t) \mid X, L_{t-1}, Y_{t-1}\ \forall\ a_t$. This implies that the effect of A_t on Y^*_t is unconfounded conditional on X, L_{t-1}, and Y_{t-1} by the potential outcomes definition of confounding. Since researchers can control for X, L_{t-1}, and Y_{t-1} in a covariate-adjustment model, this demonstrates that covariate-adjustment models can produce an unbiased estimate of the treatment effect of a non-randomized educational intervention under assumption 1.

(Please insert Figure 6 here)

Estimates from covariate-adjustment models can be biased under violations of assumption 1. The dotted blue lines in Figure 7 show causal relationships that violate assumption 1. Suppose, for example, that the edge between M and A_t exists (i.e., the probability of treatment is a function of an unobserved, time-invariant variable). Then A_t and $Y_t(a_t)$ are d-connected conditional on X, L_{t-1}, and Y_{t-1}, so $A_t \not\perp Y^*_t(a_t) \mid X, L_{t-1}, Y_{t-1}\ \forall\ a_t$ and the effect of A_t on Y_t is confounded conditional on X, L_{t-1}, and Y_{t-1}. This conclusion is surely intuitive to most researchers, but we believe our framework is novel and useful because it provides both a graphical and causal interpretation of the assumptions necessary for methods like covariate-adjustment models to "work".

(Please insert Figure 7 here)

Conclusions:
When researchers must use observational data to evaluate a non-randomized educational intervention, the appropriate estimation method depends entirely on the assumptions that the researcher is willing to make. Because of this reality, it is imperative for researchers to clearly communicate their assumptions to the reader or audience (not to mention to themselves!) A central goal of this paper is to illustrate the utility of graphs, and single world intervention templates (SWITs) in particular, as a means of communicating these assumptions. As we illustrate in the previous section, SWITs allow researchers not only to clarify the assumptions they have made, but connect these assumptions to the potential outcomes definition of confounding to illustrate how these assumptions lead to an unbiased estimate of the treatment effect. Our perspective is that if the education research community incorporates SWITs into discussions about assumptions and causality, it could bring much-needed clarity about the assumptions underlying much of the research in the education literature.

Appendices

Appendix A. References

Imbens, G. M. and Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation. Technical report, National Bureau of Economic Research.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Nye, B., Hedges, L. V., and Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, 37(1):123–151.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. University of Washington, CSSS Working Paper No. 128.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Winship, C. and Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, 659–706.

Appendix B. Tables and Figures

Definition 1. A directed acyclic graph (DAG) is a graph containing directed edges ($\rightarrow$) and no directed cycles ($V \rightarrow \cdots \rightarrow V$).
Definition 2. The parents of a vertex V with respect to a graph G are $pa_G(V) = \{X \mid X \rightarrow V\}$.
Definition 3. A vertex A is an ancestor of a vertex D if one of two conditions holds: (a) there is a directed path $A \rightarrow \cdots \rightarrow D$ from A to D; or (b) $A = D$.
Definition 4. A distribution P is Markov with respect to a DAG G if $P(\vec{V}) = \prod_{V \in \vec{V}} P(V \mid pa_G(V))$
Definition 5. A path π between vertices X and Y consists of a sequence of distinct vertices that are connected by edges.
Definition 6. A non-endpoint vertex V on a path π is a collider if π takes the form $X \cdots \rightarrow V \leftarrow \cdots Y$. Non-endpoints that are not colliders are called non-colliders.
Definition 7. A path π d-connects vertices X and Y conditional on a set $\boldsymbol{Z}$ if X and Y are the endpoints of π, every non-collider on π is not in $\boldsymbol{Z}$, and every collier on π is an ancestor of $\boldsymbol{Z}$ (or is in $\boldsymbol{Z}$). If there is no path d-connecting X and Y given $\boldsymbol{Z}$, then X and Y are d-separated given $\boldsymbol{Z}$.
Theorem. In any distribution P that is Markov with respect to G, if X and Y are d-separated given $\boldsymbol{Z}$ then $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$ in P.

Table 1. Relevant definitions and theorems from DAG literature.

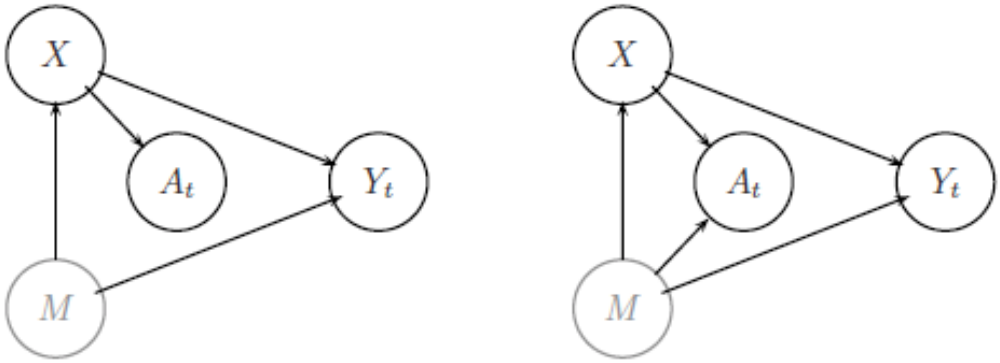
Y_t = observed test score in year *t*
A_t = indicator for whether student received the intervention in year *t*
X = observed, time-invariant characteristics (e.g., race and gender)
$\mathbf{L_t}$ = observed, time-variant characteristics (e.g., teacher and school) in year *t*
M = unobserved, time-invariant characteristics (e.g., ability)
$\mathbf{U_t}$ = unobserved, time-variant characteristics (e.g., motivation) in year *t*
Y_t* = unobserved student achievement (e.g., reading skills) in year *t*

Table 2. Variables defined for each student in general framework for pre/post evaluation of an educational intervention.

(a) Unconfounded causal effect

(b) Confounded causal effect

Figure 1. Two directed acyclic graphs (DAGs). In all figures, black vertices represent observed variables, while gray vertices represent unobserved variables. Black edges represent causal relationships between variables, while the red dotted line is the treatment effect of interest.



(a) $A_t \perp\!\!\!\perp Y_t|X$

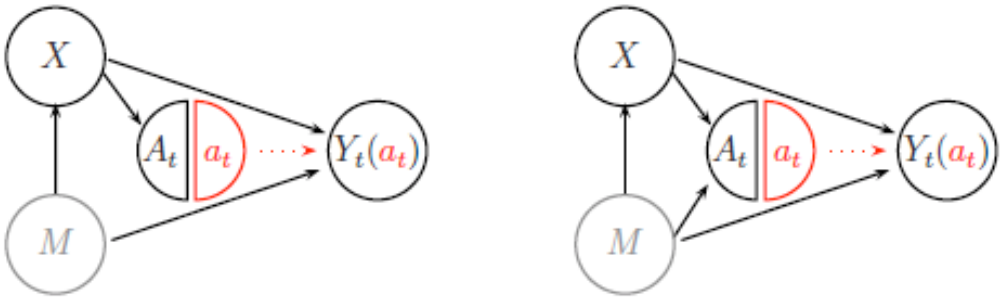
(b) $A_t \not\perp\!\!\!\perp Y_t|X$

Figure 2. Two DAGs illustrating (a) conditional independence and (b) conditional dependence of A_t and Y_t given X.

(a) Unconfounded causal effect

(b) Confounded causal effect

Figure 3. Two single-world intervention templates (SWITs).



(a) Unconfounded causal effect conditional on X

(b) Confounded causal effect conditional on X

Figure 4. Two SWITs illustrating (a) an unconfounded causal effect of A_t on Y_t given X and (b) a confounded causal effect of A_t on Y_t given X.

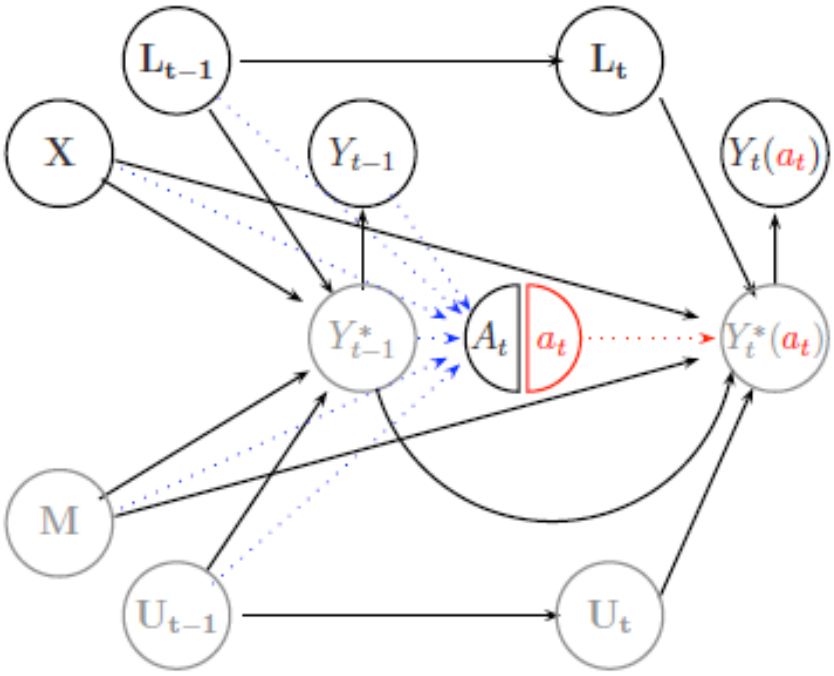


Figure 5. General SWIT for pre/post evaluation of educational intervention (see Table 2 for a definition of each variable).

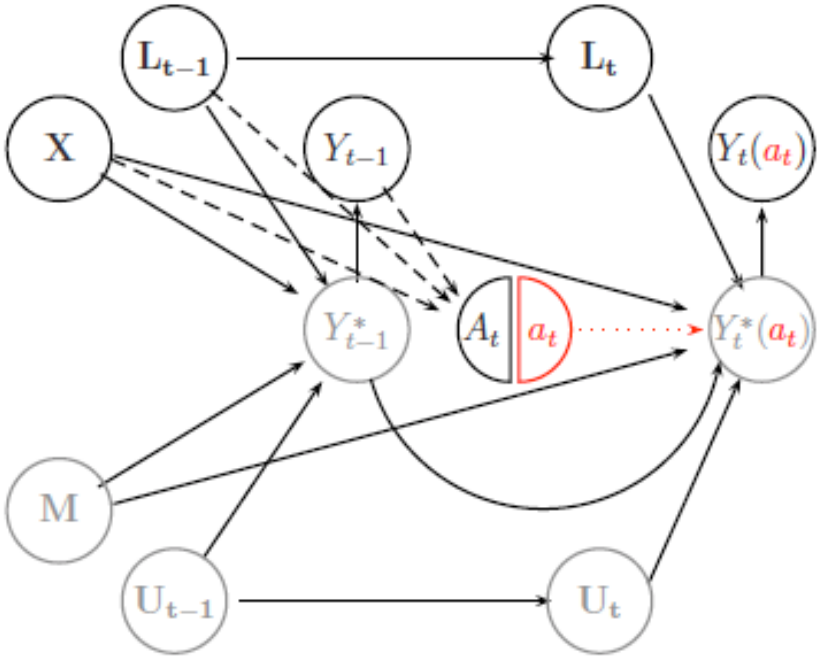


Figure 6. General SWIT under assumptions that justify a covariate-adjustment model. Dashed black lines represent edges that are allowed to exist for a covariate-adjustment model to produce an unbiased estimate of the treatment effect.

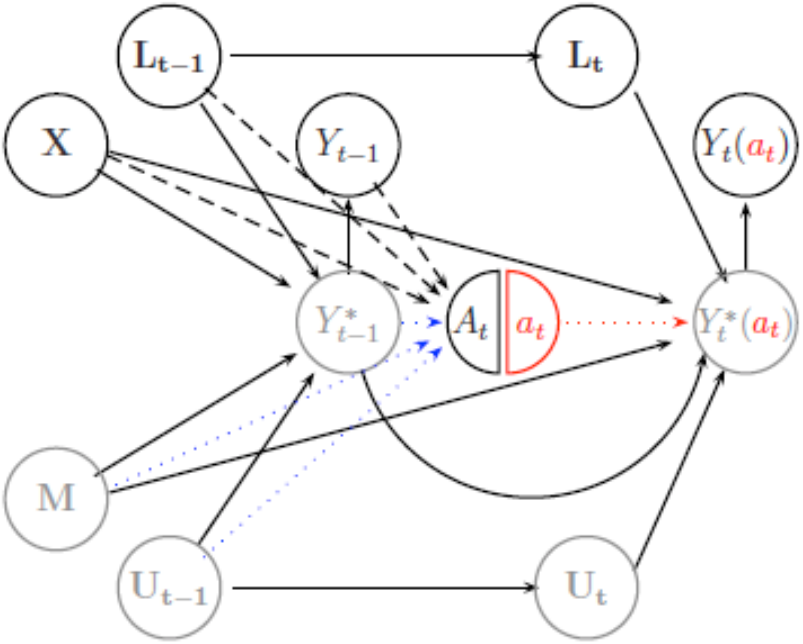


Figure 7. General SWIT illustrating violations of the assumptions that justify a covariate-adjustment model. Dotted blue lines represent edges that cannot exist for a covariate-adjustment model to produce an unbiased estimate of the treatment effect.